This is the appendix to the paper "Preserving Inversion Phylogeny Reconstruction" by Matthias Bernt, Kun-Mao Chao, Jyun-Wei Kao, Martin Middendorf, and Eric Tannier published in the Proceedings of WABI 2012.

Proof of Proposition 1

Proof. Changing the reference permutation to π has the effect of inverting the 2-sign of a node iff its first component is -1. If the 2-signs of a node and its parent are equal or differ in both positions in $\mathcal{T}^{\lambda}((\pi,\sigma),\Pi)$, then they are equal in $\mathcal{T}^{\pi}((\pi,\sigma),\Pi)$. Otherwise, *i.e.* if they differ in one position in $\mathcal{T}^{\lambda}((\pi,\sigma),\Pi)$, they differ in the second position in $\mathcal{T}^{\pi}((\pi,\sigma),\Pi)$. Thus the signs of a linear node and its parent as defined in [2], *i.e.* the second component of the 2-sign in $\mathcal{T}^{\pi}((\pi,\sigma),\Pi)$, differ iff there is one sign difference in $\mathcal{T}^{\pi}((\pi,\sigma),\Pi)$. □

Proof of Proposition 2

Proof. Let Π be a sequence of signed permutations, U and V be a linear node and its linear parent in the strong interval tree $\mathcal{T}(\Pi)$. Let π and σ be two permutations that are consistent with Π. Let c be the state of character U in $\mathcal{T}((\pi),\Pi)$ and d be the state of character U in $\mathcal{T}((\sigma),\Pi)$.

In the following we show that $c \neq d$ iff U is in any sorting inversion scenario from π to σ that is preserving for Π.

Consider the 2-signs s_U and s_V of the nodes U and V in $\mathcal{T}(\{\pi,\sigma\},\Pi)$. Then c represents the differences of the first component of s_U and s_V and d represents the differences in the second component. Thus, if $s_U = (s_U(1), s_V(2))$, then $s_V = (c \cdot s_U(1), d \cdot s_U(2))$. Then the following equivalences hold:

a) For the case $c = d$ holds i) $c = d = 1$ iff $s_U = s_V$ and ii) $c = d = -1$ iff $s_U = -s_V$.
b) For the case $c \neq d$ holds i) $c = +1$ and $d = -1$ iff $s_U(1) = s_V(1) \wedge s_U(2) \neq s_V(2)$ and ii) $c = -1$ and $d = +1$ iff $s_U(1) \neq s_V(1) \wedge s_U(2) = s_V(2)$.

Thus, $c \neq d$ iff s_U and s_V differ in exactly one position. By Proposition 1 the result follows. □

Proof of Theorem 1

Proof. In the first part of the proof, it is shown that the assignment of the permutations to the inner nodes computed with Algorithm 1 is an exact solution of the small phylogeny problem. The second part established the run time of the algorithm.

For all (linear) non root nodes U in $\mathcal{T}(\Pi)$, let R_U denote a minimum assignment of characters states for the binary character c_U defined by U with respect to its linear parent. Assume that there is an assignment Q of consistent permutations to P with score $S(Q,P) < \sum_{U \in \mathcal{T}(\Pi)} S(R_U, P)$. By Proposition 2 an assignment of character states to R'_U for each of the binary characters c_U can be given with $\sum_{U \in \mathcal{T}(\Pi)} S(R'_U, P) < \sum_{U \in \mathcal{T}(\Pi)} S(R_U, P)$. This contradicts the minimality of the assignments R_U.

The signed strong interval tree $\mathcal{T}(\Pi)$ can be computed in time $\mathcal{O}(kn)$ [3]. The characters are derived from $\mathcal{T}(\Pi)$ by comparing k-signs for each of the $\mathcal{O}(n)$ nodes. The small phylogeny problem for a single character needs time $\mathcal{O}(k)$ [4].

Since there is one character per node of $\mathcal{T}(\Pi)$, solving the small phylogeny problem for all characters needs time $\mathcal{O}(kn)$. Reconstructing the permutations for each node of the phylogeny from the character state combinations can be done in time $\mathcal{O}(kn)$. Thus, each of the four steps of the algorithm needs time $\mathcal{O}(kn)$. □

Proof of Theorem 2

Proof. A solution of the parsimony problem for preserving inversions for input permutations with linear SIT, *i.e.* a phylogeny P and a reconstruction R, can be verified in polynomial-time. The verification is to compute the score $S(P, R)$. Since $\mathcal{T}((R(u), R(v)), \Pi)$ is linear for each pair of nodes $u, v \in V(P)$, the preserving inversion distances along the edges of the tree can be computed in linear-time.

Let $C = \{c_1, \ldots, c_l\}$ be a set of binary characters, each defining the character states $+1$ or -1 for k species. Construct a tree structure with $l + 1$ nodes which is considered as linear SIT in the following. Without loss of generality the characters are assigned to the non-root nodes in pre-order. Starting with $\{+1\}^k$ as the k-sign of the root node, the k-signs of all the nodes are the result of the component-wise product of the parents k-sign with the k-states of the character. This gives a k-signed SIT which uniquely determines a set of k permutations. By Proposition 2 a solution of the preserving inversion problem corresponds to a solution of the large parsimony problem for C. □

Preserving Inversion Reconstruction for Circular gene orders Circular gene orders are also represented as signed permutations π, *i.e.* in a linear fashion, but each inversion of the complete gene order and all circular shifts are considered equivalent. Thus, the definition of *interval* needs to be extended such that element sets that are consecutive in at least one of the equivalent gene orders are considered as intervals. Note that, for a permutation with n elements and an interval I also the *complementary interval* $\{1, \ldots, n\} \setminus I$ is an interval in π. Also for circular gene orders inversions are defined as intervals. Note that, the results of complementary inversions are equivalent.

Let $\Pi^c = \{\pi_1^c, \ldots, \pi_k^c\}$ be a set of circular gene orders. By inverting the complete permutation and applying circular shift, there exist linear oriented gene orders $\Pi^l = \{\pi_1^l, \ldots, \pi_k^l\}$ with $\pi_i^l = \pi_i^c, i \in [1 : k]$, such that the first element of π_i^l is w.l.o.g. 1, i.e. $\forall_{i,j \in [1:k]} : \pi_i^l(1) = \pi_j^l(1)$.

Let L be the multiset of preserving inversions in a minimum reconstruction for the linear gene orders Π^l. Assume that a smaller multiset C of preserving circular inversions exist that is a minimum reconstruction for the circular gene orders in Π^c. From C we can construct a scenario C' of same size where each inversion that includes element 1 is replaced by its complement. Hence, for Π^l exists a preserving scenario C' with the same score as C such that each inversion is an interval in the linear directed gene orders. This contradicts the minimality of L.

The set of common intervals of a linear chromosomes is a subset of the common intervals if the same chromosomes are considered as circular. That is

for each circular interval the interval or its complement is included (or both). Thus, by preserving the intervals of the linear version also the intervals of the circular version are preserved.

Therefore, the circular undirected case can be solved by shifting and inverting the input permutations Note that it is not necessary to modify the handling of the root node because the strong interval tree for the shifted and inverted permutations has a linear root node with k-sign $\{+\}^k$.

Gamma Proteobacteria Data Set In the γ-Proteobacteria data set presented in [1] a maximum subset was determined that has linear nodes only in the corresponding SIT. The set contains four gene orders from *Escherichia coli* (eco, ecs, ece, ecc), *Salmonella enterica* subsp. *enterica serovar* Typhi str. CT18 (sty), *Salmonella typhimurium* (stm), and *Salmonella enterica* subsp. *enterica serovar* Typhi Ty2 (stt).

Since the *E. coli* gene orders are equal the data set contains four unique gene orders. If a consecutive stretch or its inverse can be found in all four gene orders it is replaced by a unique identifier. The replacements are as follows:

1 ← 1 2 -3 4 5 6 7 8 -9 -10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 -26 -27 28 29 30 31 32 33 34 35 -36
2 ← -37 38 39 40 41 42 -43 44 45 46 47 48 49 -50 -51 -52 -53 54 55 56 -57 -58 59 60 61 62 -63 64 65 66 67 68 69 70 -71 -72
3 ← -73 -74 -75 -76 77 78
4 ← 79 -80 81 -82 -83 -84
5 ← -85 86 -87 88 89 90 -91 -92 -93 -94 -95 96 -97 -98 -99 -100 101 -102 -103 -104 -105 106
6 ← -107 -108 -109 -110 -111 -112 -113 114 115 -116 -117 -118 -119 -120 -121 -122 -123 -124 125 -126 -127 128 129 130 131 132 -133 -134 -135 -136 -137 -138 -139 -140 -141 -142 -143 -144 -145 -146 -147 -148 -149
7 ← -150 151 152 -153 -154 -155 -156 -157 -158 -159 -160 -161 -162 -163 -164 -165 -166 -167 -168 -169 -170 -171 -172 -173 -174 -175 -176 -177 -178 -179 -180 -181 -182 -183 -184 -185 186 -187 -188 189 -190 -191 -192 -193 -194 195 -196 -197 198 199 200 201 -202 -203 -204 -205 -206 -207 -208 -209 -210 -211
8 ← 212 213 214
9 ← 215 -216 217 218 219 220 221 222 223 224 225 226 -227
10 ← 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 -243 -244

The final data set in fasta format.

```
>eco,ecs,ece,ecc
1   2   3   4   5   6   7   8   9  10
>sty
1   2   4  -3   5   6  -8  -9  -7  10
>stm
1   2  -4  -3   5   6   7   8   9  10
>stt
1  -5   3  -4  -2   6  -8  -9  -7  10
```

As phylogeny the corresponding subtree of the breakpoint distance based phylogeny presented in [1] was used. The inversion distance based phylogeny is identical with respect to the seven species but includes two other species. The maximum likelihood tree could not be used, since it contains a multifurcation at the subtree of interest.

Burkholderia Data Set In the following the data sets are given as fasta file. Markers representing more than one gene are listed.

Chromosome

```
>ambifaria_ammd,ambifaria_mc40-6,cenocepacia_hi2424,
cenocepacia_mc0-3,sp._383,vietnamiensis_g4
16 17 30 45 47 52 74 81
>cenocepacia_au_1054
16 -17 30 45 47 52 74 81
>cenocepacia_j2315
16 -17 30 -74 -52 -47 -45 81
>pseudomallei_1106a,pseudomallei_1710b,pseudomallei_668
16 17 30 45 -52 -47 74 81
>pseudomallei_k96243
16 17 30 45 52 -47 74 81
```

17 ← 17 18 19 20 21 22 23 24 25 26 27 28
30 ← 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
45 ← 45 46
47 ← 47 49 50 51
52 ← 52 53 54 55
74 ← 74 75 76 77 78 79 80
81 ← 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98

Plasmid

```
>ambifaria_ammd,ambifaria_mc40-6,cenocepacia_au_1054,
cenocepacia_hi2424,cenocepacia_j2315,cenocepacia_mc0-3,sp._383
14 1 2 4 5 8 11
>pseudomallei_1106a,pseudomallei_1710b,pseudomallei_668,
pseudomallei_k96243
14 -2 -1 4 -8 5 11
>thailandensis_e264
14 1 2 4 -8 5 11
>vietnamiensis_g4
14 -1 2 4 5 8 11
```

2 ← 2 3
8 ← 8 9 10
11 ← 11 12 13

Chromosome Plasmid

```
>ambifaria_ammd,ambifaria_mc40-6,cenocepacia_au_1054,cenocepacia_hi2424,
cenocepacia_j2315,cenocepacia_mc0-3,mallei_nctc_10247,mallei_savp1,
pseudomallei_1106a,pseudomallei_1710b,pseudomallei_668,sp._383,
thailandensis_e264,vietnamiensis_g4
62 66 71
>mallei_atcc_23344
-66 -62 71
>pseudomallei_k96243
-62 66 71
```

$62 \leftarrow 62\ 63\ 64\ 65$
$66 \leftarrow 66\ 67\ 68\ 69\ 70$

References

1. E. Belda, A. Moya, and F. J. Silva. Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. *Mol Biol Evol*, 22:1456–1467, 2005.
2. S. Bérard, A. Bergeron, C. Chauve, and C. Paul. Perfect sorting by reversals is not always difficult. *IEEE/ACM Trans Comput Biol Bioinf*, 4:4–16, 2007.
3. A. Bergeron, C. Chauve, F. de Montgolfier, and M. Raffinot. Computing common intervals of k permutations, with applications to modular decomposition of graphs. *SIAM J Discrete Math*, 22:1022–1039, 2008.
4. W. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst Zool*, 20:406–416, 1971.